A Survey on Bias and Fairness in Machine Learning

Year: 2019 | Citations: 5106 | Authors: Ninareh Mehrabi, Fred Morstatter, N. Saxena, Kristina Lerman, A. Galstyan

---

Abstract

With the widespread use of artificial intelligence (AI) systems and applications in our everyday lives, accounting for fairness has gained significant importance in designing and engineering of such systems. AI systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations. More recently some work has been developed in traditional machine learning and deep learning that address such challenges in different subdomains. With the commercialization of these systems, researchers are becoming more aware of the biases that these applications can contain and are attempting to address them. In this survey, we investigated different real-world applications that have shown biases in various ways, and we listed different sources of biases that can affect AI applications. We then created a taxonomy for fairness definitions that machine learning researchers have defined to avoid the existing bias in AI systems. In addition to that, we examined different domains and subdomains in AI showing what researchers have observed with regard to unfair outcomes in the state-of-the-art methods and ways they have tried to address them. There are still many future directions and solutions that can be taken to mitigate the problem of bias in AI systems. We are hoping that this survey will motivate researchers to tackle these issues in the near future by observing existing work in their respective fields.